Vietnam Journal of Science and Technology ***57*** *(5) (2019) 617-630*

doi:10.15625/2525-2518/57/5/13751

SCHEDULING FOR MASSIVE MIMO USING CHANNEL AGING UNDER QoS CONSTRAINTS

Hung Pham[1, *], Bac Dang Hoai[2], Ban Nguyen Tien[2]

[1]*The IT faculty of NAEM, 31 Phan Dinh Giot Street, Ha Noi*

[2]*Posts and Telecommunications Institute of Technology (PTIT), Nguyen Trai street, Ha Noi*

[*]Email: *hungp@niem.edu.vn*

Received: 11 April 2019; Accepted for publication: 22 July 2019

Abstract. Massive multiple-input multiple-output (MIMO) networks support QoS (Quality of Service) by adding a new sublayer Service Data Adaption Protocol on the top of Packet Data Convergence Protocol layer to map between QoS flows and data radio bearers. In downlink for Guaranteed Bit Rate (GBR) flows, the gNB guarantees the Guaranteed Flow Bit Rate (GFBR) that defines the minimum bit rate the QoS flow can provide. So, one of the most important requirements is the minimum rate. The channel aging helps to improve the sum-rate of Massive MIMO systems by serving more users to increase the spatial multiplexing gain without incurring additional pilot overhead. In this paper, a novel scheduler, termed QoS-Aware scheduling, is designed and proposed for Massive MIMO to use the channel aging to increase the sum-rate but guarantee the minimum bit rate per user to support QoS. We investigate how many users are enough to serve to maximize the sum-rate while keeping the data rate per user meeting a given threshold. Through the numerical analysis we confirmed that QoS-Aware scheduling can guarantee a minimum rate per user and get a higher useful through-put (goodput) than conventional channel aging schedulers.

Keywords: Massive MIMO, scheduling, QoS, channel aging.

Classification numbers: 4.3.1, 4.3.3.

1. INTRODUCTION

Massive multiple-input multiple-output (MIMO) is a key technique to achieve high throughput for 5G networks. The idea of Massive MIMO is using a very large number of antennas at the base station (BS) to serve multi-users (MU-MIMO) with few antennas per one. Normally, each base station has hundreds of antennas providing the connection for tens of single-antenna users. The optimal performance can be achieved when a coding scheme called dirty paper code (DPC) is used [1]. The throughput of the Massive MIMO system increases with the number of antennas at the BS and users [2]. Furthermore, the optimal performance can be nearly achieved by using linear precoders of maximum ratio transmission (MRT) or zero-forcing (ZF) when the number of antennas at the BS is very large [3], [4]. Although the advantage of Massive MIMO is clear, it still has some challenges to solve to get the optimal performance.

In order to increase the throughput of Massive MIMO systems, it is required for the BS to acquire an accurate channel state information (CSI) for precoding before transmission. In frequency division duplex (FDD) operation, the training overhead is proportional to the number of antennas, while in time division duplex (TDD), it is only scaled with the number of scheduled users. Therefore, many previous works have adopted TDD operation in Massive MIMO systems [5– 8]. In the TDD operation, the CSI is estimated in the uplink training period. The BS will select a subset of users in the cell to serve in the downlink transmission period. This scheduling process has a lot of algorithms aiming for different purposes.

To increase the data rate of users as well as maximize the throughput of the whole system, the scheduling can select users with the best channel gains and the least requirement in transmit power in each time frame to serve [9–11]. It is termed as Maximum Rate method. The issue of Maximum Rate is that the users with bad channel gains are almost never served so it is not acceptable from customer's perspective. Another way to improve the throughput is optimization of the power control, which increases the power for user with low channel gain and vice versa [12]. Another paper concerns about the total power when researching about the energy and spectral efficiency rather than to optimize the power of each user [13]. To solve the issue of Maximum Rate, Proportion Fairness takes into account users' past average data to provide an equal rate for all users. The goal of this algorithm is to maintain a fairness among users [14].

Because the number of pilot users is limited comparing to the number of antennas [15–17] to maximize the spectral efficiency so the number of scheduled users is also limited. To save resources, antenna grouping, and user grouping are considered to improve the spectral efficiency [18–20]. Authors in [21], [22] save resources by reusing pilot sequences. Another way to improve the throughput of the whole system is adding more users to serve without resources to extra CSI estimation. The current CSI of these users are estimated by the amount of channel variation of each user according to aged CSI samples. This channel aging effect in massive MIMO has received a lot of attention from researchers [23–26].

In [27], the authors proposed an Opportunistic User Scheduling using channel aging to increase the spatial multiplexing gain by serving more and more users without incurring additional pilot overhead. In 5G, physical layer in radio network will support QoS to guarantee the quality of service from the core network to end users [28]. Especially, in both downlink and uplink, the BS guarantees the Guaranteed Flow Bit Rate (GFBR) for Guaranteed Bit Rate (GBR) flows [29].

Motivated by this, we propose a novel scheduling using aged CSI, termed QoS-Aware that exploits the aged CSI to not only maximize the throughput of the system but also support the QoS by guaranteeing the minimum rate per user using GBR flow. The performance of QoS-Aware is investigated in terms of goodput (the total throughput of all users who have rates higher than the minimum rate), badput (the total throughput of all users who have rates lower than the minimum rate), the number of goodput and badput users in comparison with the scheduler in [27]. Our results show that our algorithm can guarantee the minimum rate per user and get a higher goodput than the algorithm in [27].

Notation: We use normal letters (e.g., a) for scalars, lowercase and uppercase boldface letters (e.g., $\mathbf{h}$ and $\mathbf{H}$) for column vectors and matrices. $\mathbf{I}_N$ and $\mathbf{0}_N$ are the identity matrix and all-zero matrices of size $N \times N$. For a matrix $\mathbf{A}$, $\mathbf{A}^T$ is the transpose matrix, $\mathbf{A}^*$ the conjugate transpose, and $\mathrm{tr}(\mathbf{A})$ the trace. $\mathbb{E}[\cdot]$ is the statistical expectation operator.

2. SYSTEM MODEL

We consider a single-cell multi-user massive MIMO consisting of one BS and many single antenna users. Let $\mathbb{K}_{\mathrm{a}} = \{1, 2, \cdots, K_{\mathrm{a}}\}$ be the index set of these users, where K_{a} is the total number of users. The BS is equipped with M antennas, where $M \geq K_{\mathrm{a}}$. For simplicity, it is assumed that the BS has been aware of all the users that want to be served keep unchanged in the considered frames. The system operates in TDD mode and has perfect channel reciprocity. Let T be the duration of a frame in terms of symbols. In each frame the first τ_{p} symbols are used for uplink training then the remaining ($T - \tau_{\mathrm{p}}$) symbols are used for downlink data transmission.

We assume that the channels are frequency-flat and that the channel coefficients keep constant during each frame. Let $\mathbf{g}_k[n] \in \mathbb{C}^{M\times 1}$ be the fast fading channel coefficients, which vary independently frame-by-frame. For analytical tractability, we consider Rayleigh fading channel model where the entries of $\mathbf{g}_k[n] \in \mathbb{C}^{M\times 1}$ are independently identically distributed (i.i.d.) according to the Gaussian distribution with zero mean and unit variance. Let β_k be the large-scale fading channel coefficient from user $k \in \mathbb{K}_{\mathrm{a}}$ to the BS, which does not change in the considered frames. The uplink channel coefficients from user $k \in \mathbb{K}_{\mathrm{a}}$ is determined as $\mathbf{h}_k[n] = \sqrt{\beta_k}\mathbf{g}_k[n] \in \mathbb{C}^{M\times 1}$.

In this paper, we assume that the channel coefficient vectors vary from frame to frame due to the channel aging effect. For analytical convenience, the relationship between the channel coefficient vectors in two consecutive frames is characterized by an auto-regressive model of order 1 such that [27]

$$\mathbf{h}_k[n] = \alpha_k \mathbf{h}_k[n-1] + \mathbf{e}_k[n] \tag{1}$$

where α_k is the temporal autocorrelation factor for user k, $\mathbf{h}_k[n-1]$ is the channel coefficient vector in the previous frame of user k, and $\mathbf{e}_k[n]$ is the uncorrelated channel coefficient innovation due to channel aging. In principle, α_k depends on propagation geometry, velocity of the user, and antenna characteristics [30]. For simplicity, however, it is assumed that α_k remains unchanged in the considered frames for all $k \in \mathbb{K}_{\mathrm{a}}$.

2.1. Uplink training

In the training stage of frame n, the BS selects randomly a fixed number of K_{p} users out of the K_{a} users in the round-robin manner for training purpose. Let $\mathbb{K}_{\mathrm{p}}[n] \in \mathbb{K}_{\mathrm{a}}$ be the set of indices of these selected users in frame n. After being informed, the selected users simultaneously transmit predetermined mutually orthogonal pilot sequences of length $\tau_{\mathrm{p}} \geq K_{\mathrm{p}}$ using the same transmit power of p_{p}. Note that p_{p} may vary frame-by-frame and that power control during the training stage is left for future work. Let $\mathbf{v}_k^H[n] \in \mathbb{C}^{1\times\tau_{\mathrm{p}}}$ be the pilot sequence sent by user $k \in \mathbb{K}_{\mathrm{p}}[n]$. The received training signal at the BS is

$$\mathbf{Y}_{\mathrm{r}}[n] = \sum_{k\in\mathbb{K}_{\mathrm{p}}} \sqrt{\tau_{\mathrm{p}} p_{\mathrm{p}}}\mathbf{h}_k[n]\mathbf{v}_k^H[n] + \mathbf{Z}[n] \tag{2}$$

where p_{p} is the average transmit power at each user, $\mathbf{Z}[n] \in \mathbb{C}^{M\times\tau_{\mathrm{p}}}$ is additive white Gaussian noise matrix with i.i.d. entries of $\mathcal{CN}(0, \sigma_{\mathrm{r}}^2\mathbf{I}_M)$. The MMSE estimate of $\mathbf{h}_k[n]$ is given by [31]:

$$\mathbf{h}_k[n] = \frac{\sqrt{\tau_{\mathrm{p}} p_{\mathrm{p}}}}{\sigma_{\mathrm{r}}^2 + \tau_{\mathrm{p}} p_{\mathrm{p}}} \mathbf{Y}_{\mathrm{r}}[n]\mathbf{v}_k[n]. \tag{3}$$

Moreover, due to the orthogonality principle of MMSE estimation, $\mathbf{h}_k[n]$ can be decomposed into two uncorrelated components as follows

$$\mathbf{h}_k[n] = \mathbf{h}_k[n] + \mathbf{h}_k[n] \tag{4}$$

where $\mathbf{h}_k[n]$ is the uncorrelated estimation error vector. The $\mathbf{h}_k[n]$ is a vector with i.i.d. entries $\mathcal{CN}(0, \xi_k\mathbf{I}_M), \xi_k = \dfrac{\tau_{\mathrm{p}} p_{\mathrm{p}} \beta_k^2}{\tau_{\mathrm{p}} p_{\mathrm{p}} \beta_k + \sigma_r^2}$.

2.2. Downlink Transmission

After training period, the BS selects the scheduling user set $\mathbb{K}_{\mathrm{s}}(K_{\mathrm{s}} \geq K_{\mathrm{p}})$ to transmit the data signals. Let $\mathbf{x}[n] \in \mathbb{C}^{K_{\mathrm{s}}\times 1}$ is the signal vector for K_{s} users and $\mathbb{E}\left\{\| \mathbf{x}[n] \|^2\right\} = 1$.

The BS uses a linear precoding matrix $\mathbf{F} \in \mathbb{C}^{M\times K_{\mathrm{s}}}$ which is a function of channel estimate $\mathbf{H}[n]$ to map $\mathbf{x}[n]$ to its transmit antennas. The power is allocated for k-th user is $p_k[n]$ with the power constraint $\sum_{k=1}^{K_{\mathrm{s}}} \left|\mathbf{f}_k[n]\right|^2 p_k[n] \leq P$. The received signal at the k-th user can be written as

$$\begin{aligned}
\mathbf{y}_k[n] &= \mathbf{h}_k^T[n]\mathbf{f}_k[n]\sqrt{p_k[n]}x_k[n] + \sum_{l=1,l\neq k}^{K_{\mathrm{s}}} \mathbf{h}_k^T[n]\mathbf{f}_l[n]\sqrt{p_l[n]}x_l[n] + \mathbf{n}_k \\
&= \mathbb{E}\{\mathbf{h}_k^T[n]\mathbf{f}_k[n]\}\sqrt{p_k[n]}x_k[n] + \sum_{l=1}^{K_{\mathrm{s}}} \mathbf{h}_k^T[n]\mathbf{f}_l[n]\sqrt{p_l[n]}x_l[n] \\
&\quad -\mathbb{E}\{\mathbf{h}_k^T[n]\mathbf{f}_k[n]\}\sqrt{p_k[n]}x_k[n] + \mathbf{n}_k
\end{aligned} \tag{5}$$

where $\mathbf{h}_k[n]$ is the channel vector for the k-th user, and $\mathbf{f}_k[n]$ is the k-th column of matrix $\mathbf{F}[n]$.

The instantaneous SINR for the k-th user can be written as:

$$\gamma_k[n] = \frac{p_k[n]\left|\phi_k\right|^2}{\sum_{l=1}^{K_{\mathrm{s}}} p_l[n]\mathbb{E}\{\left|\mathbf{h}_k^T[n]\mathbf{f}_l[n]\right|^2\} - p_k[n]\left|\phi_k\right|^2 + \sigma^2} \tag{6}$$

where $\phi_k = \mathbb{E}\{\mathbf{h}_k^T[n]\mathbf{f}_k[n]\}$.

The achievable rate of the k-th user is

$$\mathbf{R}_k[n] = \log_2(1 + \gamma_k[n]) \tag{7}$$

The sum-rate of the system is:

$$\mathbf{R}_{sum}[n] = \sum_{l=1}^{K_s} log_2(1+\gamma_k[n]) \tag{8}$$

2.3. Channel aging

Channel variation occurs to the remain user set $\mathbb{K}_r[n] = \mathbb{K}_a \setminus \mathbb{K}_p[n]$ due to the time difference between channel estimation and channel use in downlink period. We define the channel variation coefficient $\delta_k[n]$ which measures the channel variation between the last time slot l_k when the channel of user k -th is estimated and the current time slot n

$$\begin{aligned} \delta_k[n] &= \alpha_k^{\,n-l_k} \\ \mathbf{h}_k[n] &= \delta_k[n]\mathbf{h}_k[l_k] + \tilde{\mathbf{e}}_k[n] \end{aligned} \tag{9}$$

where, $\tilde{\mathbf{e}}_k[n] = \delta_k[n]\mathbf{h}_k[l_k] + \mathbf{e}_k[n]$ is. i.i.d with zero mean and variance $\beta_k - \delta_k^2[n]\xi_k$.

3. ACHIEVABLE SUM-RATE UNDER CHANNEL AGING

We investigate the achievable downlink sum-rate of when using aged CSI for scheduling. The current estimated CSI $\mathbf{h}_k[n]$ of user k -th is derived from the last estimated $\mathbf{h}_k[l_k]$ at time slot l_k.

$$\mathbf{h}_k[n] = \delta_k \mathbf{h}_k[l_k] \tag{10}$$

Then the received signal of user k -th is

$$\begin{aligned} \mathbf{y}_k[n] &= \mathbf{h}_k^T[n]\mathbf{F}[n]\mathbf{P}\mathbf{x}[n] + \tilde{\mathbf{e}}_k[n]\mathbf{F}[n]\mathbf{P}\mathbf{x}[n] + \mathbf{n}_k[n] \\ &= \mathbf{h}_k^T[n]\mathbf{f}_k[n]\sqrt{p_k}\,x_k[n] + \sum_{l=1,l\neq k}^{K_s} \mathbf{h}_k^T[n]\mathbf{f}_l[n]\sqrt{p_l}\,x_l[n] + \tilde{\mathbf{e}}_k[n]\mathbf{F}[n]\mathbf{P}\mathbf{x}[n] + \mathbf{n}_k \\ &= \mathbb{E}\{\mathbf{h}_k^T[n]\mathbf{f}_k[n]\}\sqrt{p_k}\,x_k[n] + \epsilon_k[n] \end{aligned} \tag{11}$$

where $\epsilon_k[n] = \sum_{l=1}^{K_s} \mathbf{h}_k^T[n]\mathbf{f}_l[n]\sqrt{p_l}\,x_l[n] - \mathbb{E}\{\mathbf{h}_k^T[n]\mathbf{f}_k[n]\}\sqrt{p_k}\,x_k[n] + \tilde{\mathbf{e}}_k[n]\mathbf{F}[n]\mathbf{P}\mathbf{x}[n] + \mathbf{n}_k[n]$

3.1. Maximum ratio transmission

The precoding matrix is given as

$$\begin{aligned} \mathbf{F}[n] &= \mathbf{H}_S^*[n] \\ \mathbb{E}\{\mathbf{h}_k^T[n]\mathbf{h}_k^*[n]\} &= \xi_k\delta_k^2[n]M \\ \mathbb{E}\{|\mathbf{h}_k^T[n]\mathbf{h}_k^*[n]|^2\} &= \xi_k^2\delta_k^4[n](M^2+M) \\ \mathbb{E}\{|\mathbf{h}_k^T[n]\mathbf{h}_l^*[n]|^2\} &= \xi_k\delta_k^2[n]\xi_l\delta_l^2[n]M \\ \mathbb{E}\{|\tilde{\mathbf{e}}_k[n]\mathbf{H}_S^*[n]\mathbf{P}\mathbf{x}[n]|^2\} &= (\beta_k - \delta_k^2[n]\xi_k) \end{aligned} \tag{12}$$

hence,

$$\gamma_k^{MRT}[n] = \frac{p_k \xi_k^2 \delta_k^4[n] M^2}{\sum_{l=1}^{K_s} \xi_k \delta_k^2[n] p_l \xi_l \delta_l^2[n] M + (\beta_k - \delta_k^2[n] \xi_k + \sigma^2)} \tag{13}$$

3.2. Zero-Forcing

For the Zero-Forcing, the precoding matrix is

$$\begin{aligned}
&\mathbf{F}[n] = \mathbf{H}_S^*[n](\mathbf{H}_S^T[n]\mathbf{H}_S^*[n])^{-1} \\
&\mathbb{E}\{\mathbf{h}_k^T[n]\mathbf{h}_k^*[n](\mathbf{h}_k^T[n]\mathbf{h}_k^*[n])^{-1}\} = 1 \\
&\mathbb{E}\{|\mathbf{h}_k^T[n]\mathbf{h}_k^*[n](\mathbf{h}_k^T[n]\mathbf{h}_k^*[n])^{-1}|^2\} = 1 + \frac{(\beta_k - \delta_k^2 \xi_k)}{\delta_k^2 \xi_k (M - K_s)} \\
&\mathbb{E}\{|\mathbf{h}_k^T[n]\mathbf{h}_l^*[n](\mathbf{h}_l^T[n]\mathbf{h}_l^*[n])^{-1}|^2\} = \frac{(\beta_k - \delta_k^2 \xi_k)}{\delta_l^2 \xi_l (M - K_s)},\ l \neq k \\
&\mathbb{E}\{|\tilde{\mathbf{e}}_k[n]\mathbf{F}[n]\mathbf{Px}[n]|^2\} = (\beta_k - \delta_k^2[n]\xi_k)
\end{aligned} \tag{14}$$

Hence,

$$\gamma_k^{zf}[n] = \frac{p_k (M - K_s)}{\sum_{l=1}^{K_s} \frac{p_l (\beta_k - \delta_k^2[n]\xi_k)}{\delta_l^2[n]\xi_l} + \varphi(M - K_s)}, \varphi = (\beta_k - \delta_k^2[n]\xi_k + \sigma^2) \tag{15}$$

3. QOS-AWARE DESIGN

Using aged CSI can increase multiplexing gain but it leads the average rate of user to go down and it may not meet the minimum rate from QoS requirement. To alleviate this problem, we propose an opportunistic user scheduling algorithm that not only schedules more users to achieve higher multiplexing gain but also guarantee the minimum rate per user to support QoS. The key idea in QoS-Aware design is checking the achievable rate for all selected users and the candidate user if adding this candidate to the scheduled group still meet the minimum rate from QoS requirement.

Let us formulate the scheduler's objective and the constraints. The BS gathers the channel state information $\mathbf{H}$, the total transmit power P and the minimum rate T for scheduled users from the QoS requirement. Based on the collected information, the scheduler maximizes the sum-rate of the whole system by selecting the best user subset $\mathbb{K}_s[n]$ from the pilot user set $\mathbb{K}_p[n]$ and aged CSI in each time frame.

$$\begin{aligned}
&\max_{\mathbb{K}_s[n] \in \mathbb{K}_a} \sum_{k=1}^{K_s} \log_2(1 + \gamma_k[n]) \\
&s.t \quad \sum_{k=1}^{K_s} |\mathbf{f}_k|^2 p_k \leq P \\
&\log_2(1 + \gamma_k[n]) \geq T
\end{aligned} \tag{16}$$

As can be seen in Figure 1, the scheduling consists of four parts: *estimation of channel variation coefficients* which helps the BS estimate the amount of channel variation for each user from the last CSI training, *pilot user selection* which selects a subset of users $\mathbb{K}_p[n]$ for pilot training, *uplink training* will update channel state information for group $\mathbb{K}_p[n]$, and *valid user selection for transmission* which allows more users not only to be scheduled with aged to increase the spatial multiplexing gain but also guarantee their minimum rates.

After every uplink training procedure and updating the channel state information for users in group $\mathbb{K}_p[n]$, the BS has to choose the valid users precisely who help to improve performance of the whole system but still keep the minimum rate per user. We denote the $\mathbb{K}_s[n]$ is the scheduled group and the $\mathbb{K}_c[n]$ is the candidate group. To determine whether user k to be scheduled or not, at timeslot n the rate R_l per user $l \in \{k \bigcup \mathbb{K}_s[n]\}$ have to be higher than the minimum rate T. If all users satisfy this QoS requirement then we check if adding the user k will help to increase the throughput or not $R^{pc}(k \bigcup \mathbb{K}_s[n]) > R_{sum}$. Lastly, if there is the best user k in $\mathbb{K}_c[n]$ meets both these conditions then the user k will be moved to group $\mathbb{K}_s[n]$:

$$\mathbb{K}_s[n] = k_{best} \bigcup \mathbb{K}_s[n]$$
$$\mathbb{K}_c[n] = \mathbb{K}_c[n] \setminus \{k_{best}\}$$

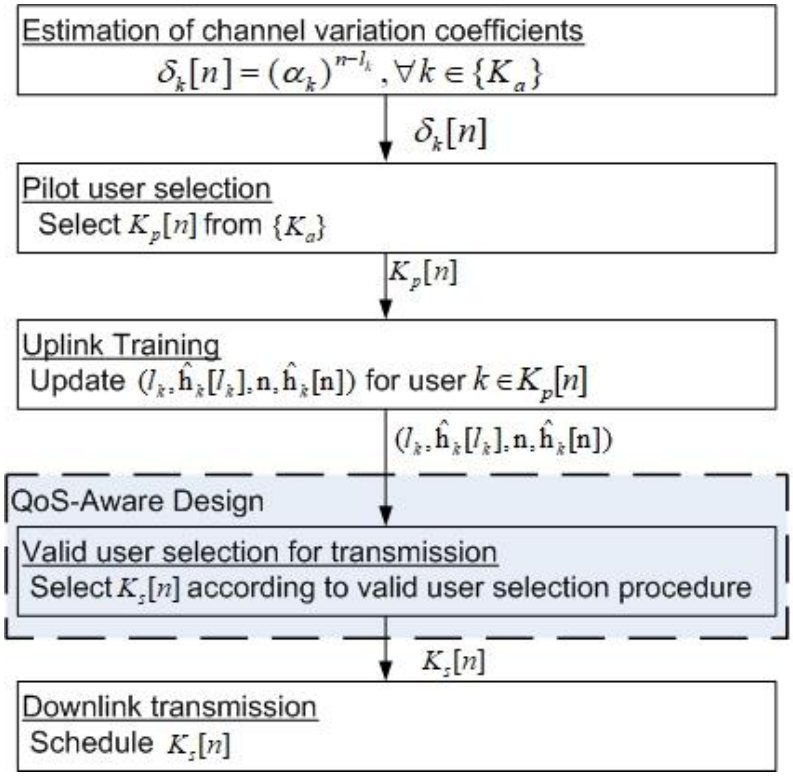

Figure 1. The overview of QoS-Aware method.

The BS will loop these procedures until there is no user meeting the conditions. These procedures are described in Algorithm 1.

Algorithm 1 valid user selection procedure

$\mathbb{K}_s[n] = \emptyset, \mathbb{K}_c[n] = \mathbb{K}_a, R_{sum} = 0$
$morevaliduser = 1$
while $morevaliduser$ **do**
  $k_{best} \leftarrow 0$
  for user $k \in \mathbb{K}_c[n]$ **do**
    $QoScheck = 1$
    for user $l \in \{k \bigcup \mathbb{K}_s[n]\}$ **do**
      Calculate $\gamma_l[n]$ follow MRT or ZF
      $R_l = \log_2(1 + \gamma_l[n])$
      if $R_l < T$ **then**
        $QoScheck = 0$
      end if
    end for
    if $(R^{pc}(k \bigcup \mathbb{K}_s[n]) > R_{sum})$ and $(QoScheck)$ **then**
      $k_{best} \leftarrow k$
      $R_{sum} \leftarrow R^{pc}(k \bigcup \mathbb{K}_s[n])$
    end if
  end for
  if $k_{best} \neq 0$ **then**
    $\mathbb{K}_s[n] = k_{best} \bigcup \mathbb{K}_s[n]$
    $\mathbb{K}_c[n] = \mathbb{K}_c[n] \setminus \{k_{best}\}$
  else
    $morevaliduser = 0$
  end if
end while

4. SIMULATION RESULTS

To measure the effect of QoS-Aware scheduling, various case studies have been done based on Massive MIMO system to compare the following scheduling policies:

- Non-QoS Scheduler (OpSac in [27]).
- QoS-Aware scheduler

We mainly compare the goodput that is the total rate of users who get the rate higher than the minimum rate T and vice versa for the badput. In all cases, we set the training sequence length $\tau_p = K_p$, and $\sigma_r = 20$ dB, $\sigma = 30$ dB.

Figure 2 compares the goodput of QoS-Aware and Non-QoS according to the number of BS antennas when $K_a = 40$ for the minimum rate $T = 0.1, 1$ and 2. All of them are using MRT precoding. It can be seen that the goodput of the system increases when the number of antennas M goes up. Moreover, the goodput of QoS-Aware is always higher the one of Non-QoS. If the T decreases than the difference of the goodput will be smaller. Especially, with

$T = 0.1$ the goodput for both QoS-Aware and Non-QoS are almost the same. Lastly, when the T decreases, the goodput of both two methods increases.

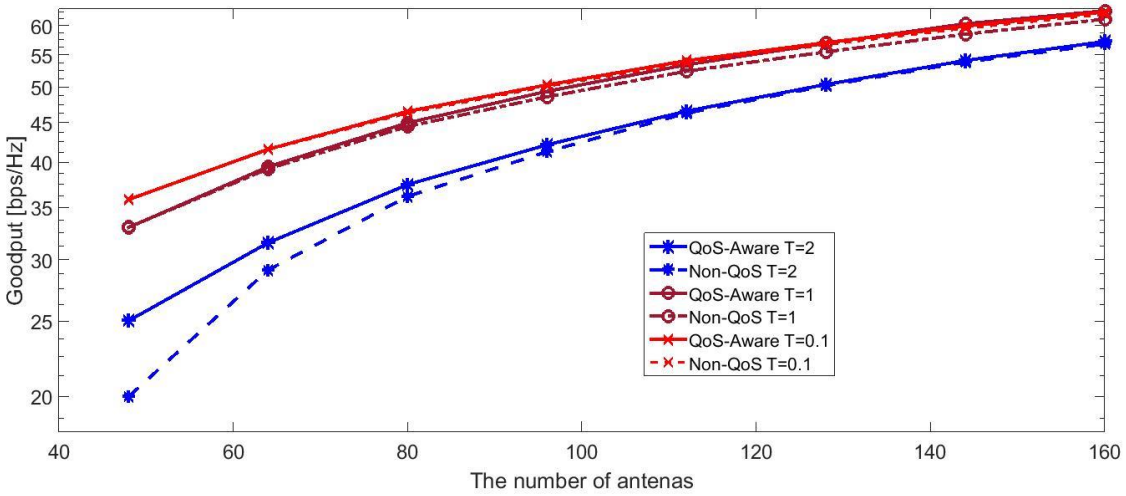

Figure 2. The comparison of goodputs when using MRT.

Figure 3 shows that the badput of QoS-Aware is less than the one of Non-QoS with the same *T*. The higher the minimum required rate *T* is, the bigger the badput is. The worst of badput is the case $T = 2$ with Non-QoS method. It is obviously that Non-QoS should not be applied for the deployment of wireless network where there are applications requiring high speed rates.

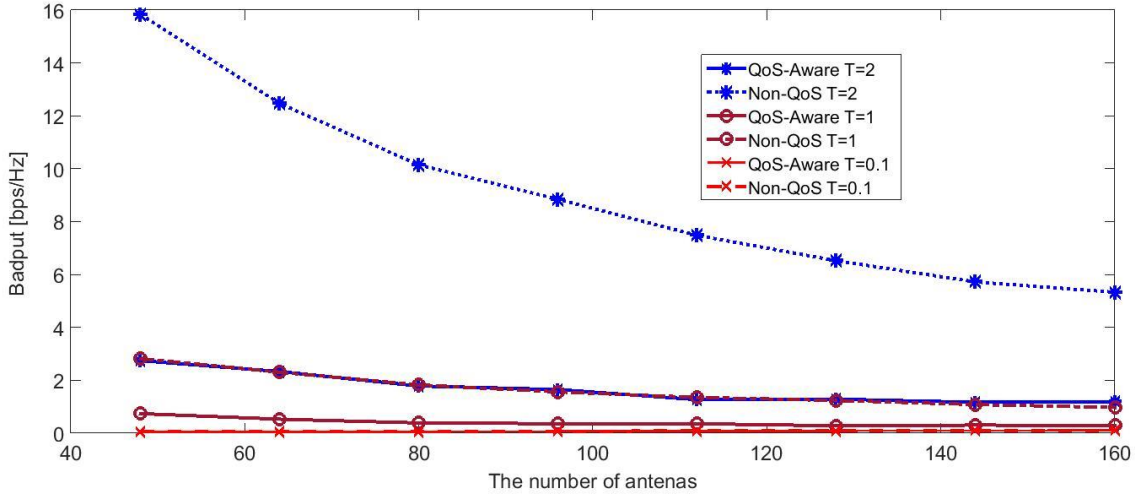

Figure 3. The comparison of badputs when using MRT.

Figure 4 shows the goodput of system according to the number of users K_a when $M = 128$. Normally, adding more users will increase the goodput of system. However, if using Non-QoS and the T is high then the goodput will go down as the case $T = 2$. It means if using the Non-QoS then serving more users can lead to most of them will have badput. This happens when the average rate of users is smaller than the minimum rate expected T. However, with

$T = 2$ it is very good for QoS-Aware that the goodput still increases even when adding more users because the scheduling will only select the best users while monitoring the total number of them to satisfy the QoS requirement.

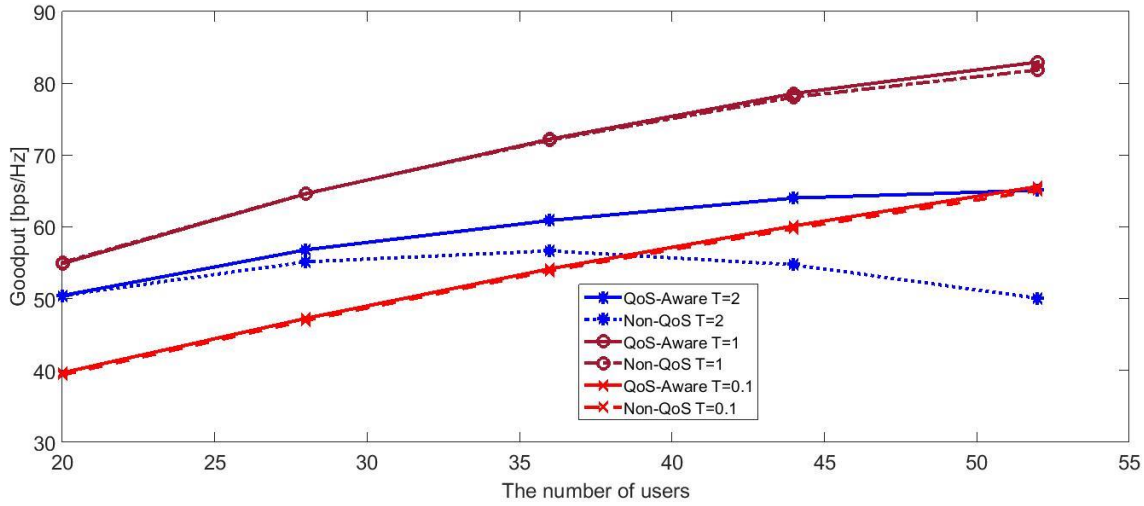

Figure 4. The goodput of system using MRT when $M = 128$.

Figure 5 shows the goodput of QoS-Aware and Non-QoS with ZF precoding when $K_a = 40$ as M increases for the minimum rate $T = 9, 10,$ and 11. It can be seen that normally the goodput of QoS-Aware are always higher than the one of Non-QoS. However, if the *T* is less than the average rate per user then the goodput of QoS-Aware will be smaller than the one of Non-QoS, for example with $T = 9$. It means if the minimum rate T is too low, the QoS-Aware scheduling is not needed. Lastly, when the number of antennas increases, the goodput of two methods goes up.

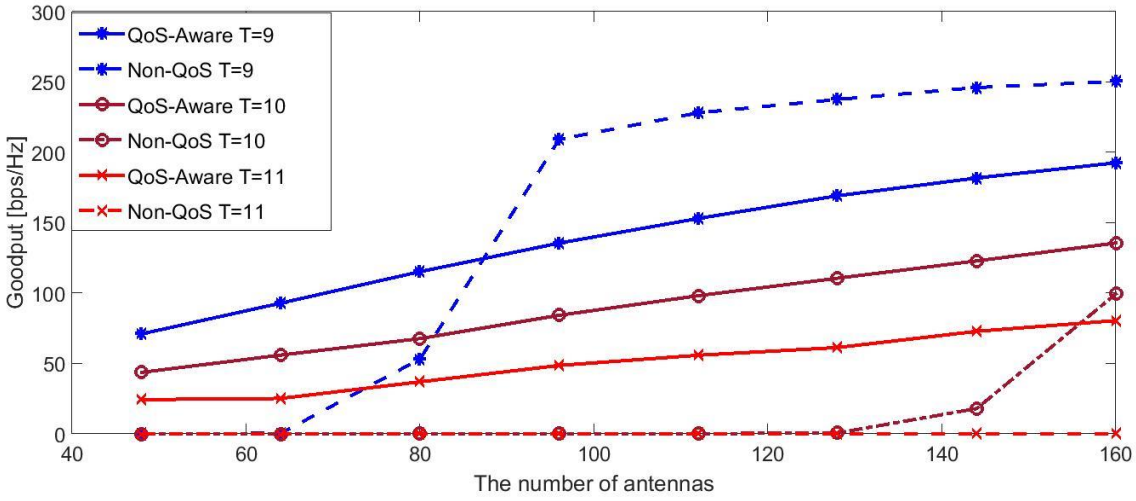

Figure 5. The comparison of goodput when using ZF with $K_a = 40$.

Figure 6 shows that the badput of both methods with ZF precoding. QoS-Aware also most has no badput because the interference is vanished when M increases. For Non-QoS, if the T is less than the average rate of users then the badput of system also goes down to zero, for example with $T = 9$. It is confirmed again if the T is low it is better not to use QoS-Aware method.

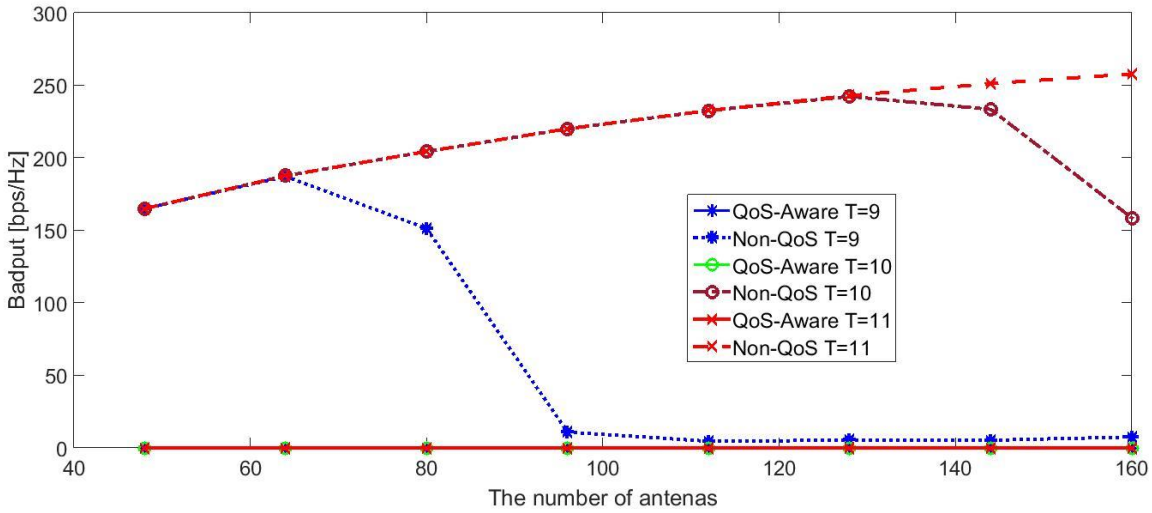

Figure 6. The comparison of badput when using ZF with $K_a = 40$.

Figure 7 show the number of scheduled users in the case of ZF precoding. As normal, the Non-Qos will serve the biggest number of users and QoS-Aware will serve fewer and fewer users if the *T* goes up. Moreover, MRT always server more users than ZF with the same *T*. For Non-QoS using MRT, almost of users will be scheduled even though many of them will have badput. On other hand, for Non-QoS using ZF only selects the users with good condition of current channel or channel aging.

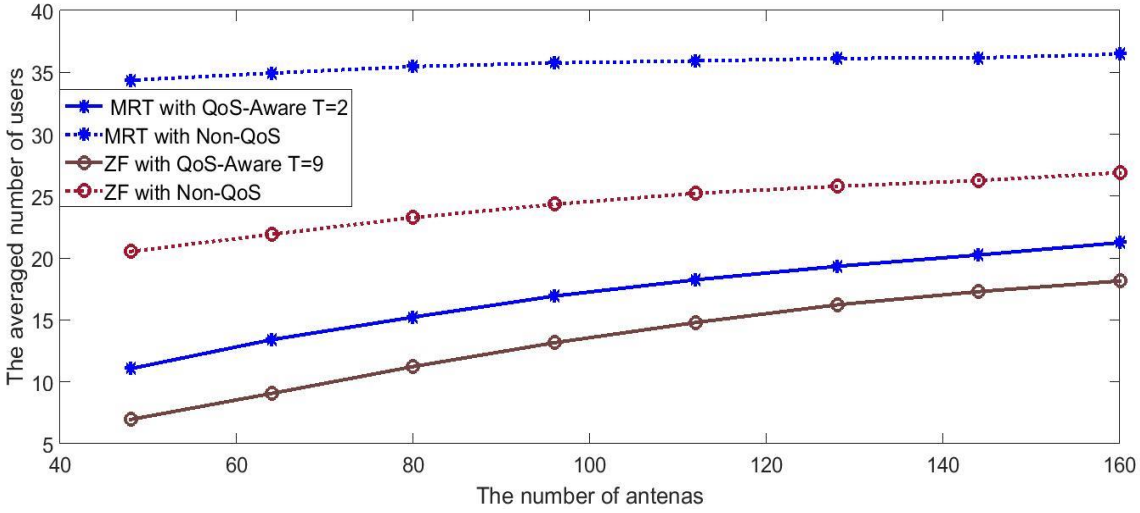

Figure 7. The number of scheduled users when using ZF and MRT.

5. CONCLUSIONS

In this paper, we proposed a novel scheduling algorithm, termed QoS-Aware, that exploits aged CSI with concern about the minimum rate per user to support QoS. We analyzed the sum rate as well as the achievable rate for the downlink when MRT and ZF are employed. According to the analytical results, we provided a scheduler that improves the sum rate by serving more users but still satisfies the minimum rate per user to guarantee the QoS. It was shown that even in the traffic jam condition for example there are too many users in the cell, the QoS-Aware will only select enough users to be scheduled to avoid the case a user experiences a low rate connection. It is really promising to deploy multi-media services efficiently on 5G wireless network.

REFERENCES

1. Costa M. - Writing on dirty paper (corresp.), IEEE Transactions on Information Theory **29** (3) (1983) 439–441.
2. Rusek F., Persson D., Lau B. K., Larsson E. G., Marzetta T. L., Edfors O., and Tufvesson F. - Scaling up mimo: Opportunities and challenges with very large arrays, IEEE Signal Processing Magazine **30** (1) (2013) 40–60.
3. Hoydis J., Brink S. ten, and Debbah M. - Massive MIMO in the ul/dl of cellular networks: How many antennas do we need?, IEEE Journal on Selected Areas in Communications **31** (2) (2013) 160–171.
4. Kong C., Zhong C., and Zhang Z. - Performance of zf precoder in downlink massive mimo with non-uniform user distribution, Journal of Communications and Networks **18** (5) (2016) 688–698.
5. Ngo H. Q., Larsson E. G., and Marzetta T. L. - Massive mu-mimo downlink tdd systems with linear precoding and downlink pilots, in 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton) (2013) 293–298.
6. Marzetta T. L. - Noncooperative cellular wireless with unlimited numbers of base station antennas, IEEE Transactions on Wireless Communications **9** (11) (2010) 3590–3600.
7. Yang H. and Marzetta T. L. - Performance of conjugate and zero-forcing beamforming in large-scale antenna systems, IEEE Journal on Selected Areas in Communications **31** (2) (2013) 172–179.
8. Larsson E. G., Edfors O., Tufvesson F., and Marzetta T. L. - Massive MIMO for next generation wireless systems, IEEE Communications Magazine **52** (2) (2014) 186–195.
9. Shariat M., Quddus A. U., Ghorashi S. A., and Tafazolli R. - Scheduling as an important cross-layer operation for emerging broadband wireless systems, IEEE Communications Surveys Tutorials **11** (2) (2009) 74–86.
10. Bohge M., Gross J., Wolisz A., and Meyer M. - Dynamic resource allocation in ofdm systems: an overview of cross-layer optimization principles and techniques, IEEE Network **21** (1) (2007) 53–59.
11. Alkhaled M., Alsusa E., and Pramudito W. - Adaptive user grouping algorithm for the downlink massive mimo systems, in IEEE Wireless Communications and Networking Conference (2016) 1–6.

12. Yoo T. and Goldsmith A. - On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming, IEEE Journal on Selected Areas in Communications **24** (3) (2006) 528–541.
13. Ngo H. Q., Larsson E. G., and Marzetta T. L. - Energy and spectral efficiency of very large multiuser mimo systems, IEEE Transactions on Communications **61** (4) (2013) 1436– 1449.
14. Afroz F., Heidery R., Shehab M., Sandrasegaran K., and Shompa S. - Comparative analysis of downlink packet scheduling algorithms in 3gpp lte networks, International Journal of Wireless Mobile Networks **7** (2015) 1–21.
15. Marzetta T. L. - How much training is required for multiuser mimo? in Fortieth Asilomar Conference on Signals, Systems and Computers (2006) 359–363.
16. Bjornson E., Larsson E. G., and Debbah M. - Massive mimo for maximal spectral efficiency: How many users and pilots should be allocated?, IEEE Transactions on Wireless Communications **15** (2) (2016) 1293–1308.
17. Hassibi B. and Hochwald B. M. - How much training is needed in multiple-antenna wireless links?, IEEE Transactions on Information Theory **49** (4) (2003) 951–963.
18. Lee B., Ngo L., and Shim B. - Antenna group selection-based user scheduling for massive mimo systems, in IEEE Global Communications Conference (2014) 3302–3307.
19. Xu Y., Yue G., Prasad N., Rangarajan S., and Mao S. - User grouping and scheduling for large scale mimo systems with two-stage precoding, in IEEE International Conference on Communications (ICC) (2014) 5197–5202.
20. Benmimoune M., Driouch E., Ajib W., and Massicotte D. - Joint transmit antenna selection and user scheduling for massive mimo systems, in IEEE Wireless Communications and Networking Conference (WCNC) (2015) 381–386.
21. You L., Gao X., Xia X., Ma N., and Peng Y. - Pilot reuse for massive mimo transmission over spatially correlated rayleigh fading channels, IEEE Transactions on Wireless Communications **14** (6) (2015) 3352–3366.
22. Sohn J., Yoon S. W., and Moon J. - On reusing pilots among interfering cells in massive mimo, IEEE Transactions on Wireless Communications **16** (12) (2017) 8092–8104.
23. Truong K. T. and Heath R. W. - Effects of channel aging in massive mimo systems, Journal of Communications and Networks **15** (4) (2013) 338–351.
24. Papazafeiropoulos A. K. and Ratnarajah T. - Deterministic equivalent performance analysis of time-varying massive mimo systems, IEEE Transactions on Wireless Communications **14** (10) (2015) 5795–5809.
25. Kong C., Zhong C., Papazafeiropoulos A. K., Matthaiou M., and Zhang Z. - Effect of channel aging on the sum rate of uplink massive mimo systems, in IEEE International Symposium on Information Theory (ISIT) (2015) 1222–1226.
26. Kong C., Zhong C., Papazafeiropoulos A., Matthaiou M., and Zhang Z. - Sum-rate and power scaling of massive mimo systems with channel aging, IEEE Transactions on Communications **63** (10) 2015.
27. Lee H., Park S., and Bahk S. - Enhancing spectral efficiency using aged csi in massive mimo systems, in IEEE Global Communications Conference (GLOBECOM) 2016, pp. 1–6.

28. 3GPP TR 38.804, Study on New Radio Access Technology, Radio Interface Protocol Aspects, Std., Mar. V14.0.0, (2017).
29. 3GPP TS 38.300, ~- 3rd Generation Partnership Project, Technical Specification Group Radio Access Network, Std., Jun. V15.2.0 (2018).
30. Jakes W. C. - Microwave mobile communications.New York: Wiley, 1974.
31. Hampton J. R. - Introduction to MIMO Communications. Cambridge University Press, 2015.